2025 ■ Vol. 21 ■ no. 3

Exploring Memory Span and Capacity: A Replication of Martin (1978, Experiment 2)

Paola Michelle Garcia Mairena [a], Cédric Champagne [a], Vincent Poulin [a], Laurie Rochon [a], Gabrielle Tao [a], Janelle Allanach [a], Osvaldo Arenas [a], Beril Arpaci [a], Nada Attrassi [a], Zineb Berrada [a], Camy Bérubé [a], Chloé Bilodeau [a], Eva Bouity [a], Audrey Boutin [a], Marika Castonguay [a], Marie Pier Cléroux [a], Émika Easton [a], Salma Errachoud [a], Rebecca Forbes [a], Arpine Gasparyan [a], Omar Goncci [a], Sophonia Jacques [a], Kenza Jarmoune [a], Ariane Jeaurond [a], Adidja Kamikazi [a], Adèle Labonté [a], Zacharie LeBrun-LeMay [a], Sabrina Maurice [a], Edith Mbayo [a], Sabrina Morris [a], Emili Osver [a], Kenza Rakhdoune [a], Emma-Line Rozon [a], Prisca Salava [a], River Seraphin [a], Sunnie Spratt [a], Alex St-Jean [a] & Idir Zaher [a]

[a]School of Psychology, University of Ottawa, Ottawa, Canada

Abstract ■ The digit span task has long been a vital component in assessing working memory in both clinical and research contexts. Martin's (1978) two-experiment research established that digit span performance reflects order-information retention rather than mere memory capacity. The following study aimed to replicate Martin's second study methodology using a larger and more diverse sample to enhance generalizability, employing modernized testing procedures. English and French participants ($n = 79$) between 17 and 74 years of age, with varying levels of education and backgrounds, were asked to complete a 45- to 60-minute study, which included order recall and digit span tasks. While our results generally support the hypothesis that digit span is a predictor of memory performance, they only partially align with Martin's original findings. We observed significant but weaker correlations between digit span scores and strict recall scores ($r = 0.32$, $p = 0.004$) compared to Martin's original report ($r = 0.63$, $p = 0.01$). Conversely, we found a significant increase in correlation with lenient recall scores ($r = 0.34$, $p = 0.002$), indicating a shift in performance patterns. Notably, our participants exhibited nearly a 50% decline in average performance compared to those in the original study, suggesting potential external influences on outcomes. Future research should consider integrating a verbal span test to simultaneously assess item and order memory and explore how factors such as age and demographic variables influence performance. This approach would provide insights into generational effects and contextual influences on memory processing.

Keywords ■ Working memory, Replication, Order Recall, Digit Span, Memory Performance.

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ The author of the target article did not respond to the invitation to review

pgarc090@uottawa.ca

10.20982/tqmp.21.3.p138

Introduction

The digit span task, a cornerstone of working memory assessment since the late 19th century (Blankenship, 1938), requires participants to recall sequences of auditorily presented digits (Weiss et al., 2016). As one of the most extensively studied neurocognitive tasks, performance is quantified by the maximum sequence length that can be reproduced (Schroeder et al., 2011). Its standardized inclusion in intelligence assessments like the Wechsler Scales (WAIS/WISC) and historical examinations (e.g., Binet tests) underscores its clinical and research utility for evaluating conditions such as learning disabilities, neurocognitive disorders, and stroke-related deficits (Conway et al., 2005; Geva et al., 2021, 2021; Giofrè et al., 2016). Lynch and Baudry (2015) highlight the ongoing debate on the capacity of short-term memory (STM). Indeed, Miller (1956) determined working memory capacity (WMC) to be 7 ± 2, while Cowan (2001) concluded that it was 4 ± 1. Wilhelm et al. (2013) question those measures of WMC and instead propose a model based on its limitations. While traditionally viewed as indicators of STM capacity, recent cognitive psychology research suggests that digit span tests may predominantly reflect an individual's ability to retain sequential order information rather than their overall capacity for item retention (Oberauer et al., 2018). This ongoing investigation into digit span contributes to variability in findings, highlighting the need for a consensus that can only be achieved through study replication (Maxwell et al., 2015). In psychology, the lack of replication has been increasingly recognized by the academic community, with some referring to it as a replication crisis (Braden, 2024). As a result, researchers are calling for more robust protocol publications to encourage replication efforts. Given the widespread use of digit span as a staple of intelligence testing, a thorough understanding of the memory processes it involves is essential for maintaining confidence in the task's validity. Therefore, replication of findings concerning the memory processes underlying digit span capacity is necessary.

Given that digit span requires information registration, temporary storage, focused attention, auditory discrimination and auditory rehearsal, it is widely accepted as being linked to working memory (Weiss et al., 2016). Working memory, conceptualized by Baddeley and Hitch (1974), provides a more comprehensive framework (see Figure 1) for understanding these processes. Their multicomponent model comprises a central executive (responsible for attentional allocation and cognitive control), a phonological loop (for verbal information maintenance), and a visuospatial sketchpad (for visual and spatial representation retention). Baddeley (2000) later added a fourth component, the episodic buffer, which integrates information from the other systems and long-term memory. This model explains how sequential processing and item retention operate as distinct processes within memory tasks.

The importance of understanding these distinctions extends beyond theoretical interest. Memory span differences have practical implications across multiple domains, including education, neuropsychology, and clinical assessment. Extensive experimental and meta-analytic studies (Huang-Pollock & Karalunas, 2010; Martinussen et al., 2005; Rapport et al., 2008; Willcutt et al., 2005) demonstrate that deficits in working memory capacity, particularly in the central executive function and the visuospatial sketchpad of Baddeley's model, are found in children with attention-deficit/hyperactivity disorder (ADHD). Furthermore, deficits in working memory capacity may serve as early markers for conditions such as Alzheimer's disease (Rosen et al., 2002). Additionally, social psychology research has revealed that stereotype threat can reduce working memory capacity, mediating its effects on standardized test performance (Schmader & Johns, 2003). In developmental psychology, declines in working memory capacity associated with aging are central to the general cognitive effects of aging (Hasher & Zacks, 1988). Given these theoretical and empirical developments, replicating Martin's (1978) study is both timely and necessary to reassess the validity of digit span within modern memory models.

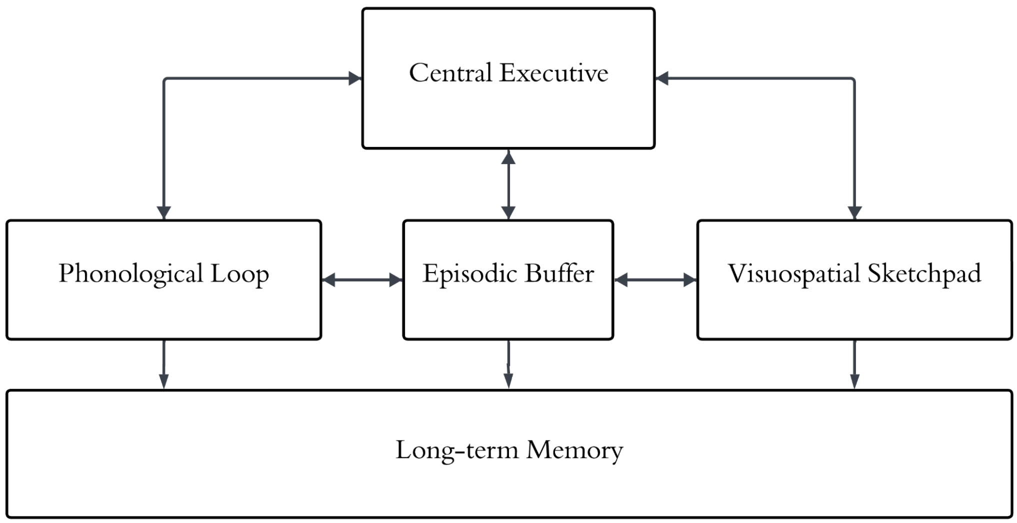
Martin's (1978) two-experiment study provided foundational evidence that digit span performance reflects order-information retention rather than general memory storage capacity. In her first experiment ($N = 38$), digit span showed correlation with order recall measures while demonstrating no significant relationship with other memory store capacities. The second experiment ($N = 16$) further isolated this relationship, revealing a strong correlation between digit span and order retention ($r = .63$, $p < .01$), but not item retention ($r = .28$). These findings suggested that digit span tasks primarily measure sequential processing abilities rather than storage capacity per se.

Subsequent research has generally supported Martin's distinction between item and order memory (Majerus et al., 2006), though questions remain about the generalizability of her original findings. The study's relatively small, homogeneous sample, drawn from the Oxford Subject Panel, and dated methodology limit the certainty with which we can apply these conclusions to broader populations or contemporary research contexts. Moreover, the increasing use of digit span tasks in clinical, educational, and neuropsychological assessments makes it imperative to establish whether they truly measure what researchers and practitioners assume they measure.

The current study aims to replicate and extend Martin's (1978) second study with two key improvements: a larger, diverse sample to enhance generalizability and modernized testing procedures. We specifically examine whether digit span performance shows a stronger correlation with order memory (assessed via strict scoring criteria) than with item memory (assessed via lenient criteria). Confirming this pattern would support the construct validity of digit span as a measure of sequential processing in working memory; null findings, however, suggest the need for theoretical refinement.

This replication serves multiple important functions. First, it addresses fundamental questions about the cognitive processes underlying digit span performance. Second, it provides an empirical test of whether mid-20th-century memory findings remain valid in contemporary research contexts. Finally, by clarifying what digit span tasks truly measure, the results may inform their appropriate use across various applied settings where they are

Figure 1 ■ Baddeley and Hitch's Working Memory Model (1974). Working memory conceptualization as formulated by Baddeley & Hitch (1974).



commonly employed as assessment tools.

In line with both the original findings and recent theoretical developments, we hypothesize that, first, digit span performance will correlate significantly with sequence order memory capacities as measured by strict scoring criteria; second, digit span performance will show limited or no significant relationship with pure item retention abilities as measured by lenient scoring criteria; and finally these differential correlations will support the interpretation that digit span tests primarily assess aspects of working memory related to sequential processing rather than general short-term storage capacity.

Through this systematic replication, we aim to contribute to both ongoing theoretical discussions about working memory architecture and the broader conversation about reproducibility in psychological science.

Methods

The methodology of this study was designed to closely follow Martin's (1978) second experiment. Key elements, such as the order of stimulus presentation and scoring methods, were preserved. However, some adjustments were made to stimuli and recall times to ensure consistency between tasks and to facilitate adaptation to an online format. Those adjustments are described in more detail in the materials and the procedure sections.

Participants

This study was conducted as part of the laboratory component of the undergraduate course Cognitive Processes (PSY 3777) at the University of Ottawa during the winter 2025 session. A non-probabilistic convenience sampling method was employed: all students enrolled in this course were asked to invite friends and family members to participate and were responsible for sending them a link to the online study, which included a unique access code. Participants had two weeks to complete the study and were instructed to reserve a 45- to 60-minute period, free from any distractions, to complete the study. Moreover, no compensation was provided to the individuals who participated in the study.

Data was obtained from 79 participants. Exclusion criteria included not completing the full study ($n = 0$) and being under 17 years of age ($n = 0$). To participate in the study, participants also had to reside in Canada and be able to understand and read English or French. Data was retained for all participants (female, $n = 37$; male, $n = 42$) aged between 17 and 74 years old ($M = 32.16$, $SD = 16.39$), from diverse ethnic backgrounds (Caucasian, African Black, Caribbean Black, East Asian, Indigenous, South Asian and West Asian), all proficient in either English or French. See Table 1 on the journal's web site for a detailed demographic profile of the participant sample.

Materials

This study was administered through an online program developed using `JavaScript` and made accessible via a website link shared with participants. This format represents a modification of Martin's (1978) original experiment, which was conducted in a laboratory setting, and was adapted here to improve accessibility for participant recruitment. The program was designed to launch in full-screen mode, minimizing on-screen distractions. It included a consent form, demographic questions, task instructions, practice sessions, stimulus presentation, an optional break and closing questions. All stimuli were displayed in the center of the screen, in bold, on a white background for optimal contrast. Due to the program's layout, participants were asked to complete the study on a computer or tablet, rather than a cell phone, to ensure proper visibility.

Stimuli

The study consisted of two tasks: a digit span task and a serial letter recall task. For the digit span task, three sets of MP3 auditory stimuli were prepared and were randomly generated by JavaScript for each participant. Each set included ten immediate digit lists made up of digits ranging from 0 to 9. Each set contained a single-length list, ranging from three to twelve digits in length. The length increased by one digit per list, reaching twelve digits by the tenth list. Lists of ten digits or less contained no repeated digits, while longer lists contained one or two repeated digits, though not consecutively. The lists were presented in ascending order, and a "ready" auditory stimulus was announced five seconds before the presentation of each sequence. The sequence was then presented at a rate of one digit per second. A standardized three-second interval was implemented between the end of each stimulus presentation and the appearance of the response box to facilitate consistent response timing and data coding. Participants had 30 seconds to report the sequence presented. This protocol was repeated 30 times.

For the order recall visual task, participants were asked to memorize a series of capital letters presented visually in the middle of their screen. We did not use predetermined lists, unlike Martin's (1978) use of Latin squares. The present study generated lists randomly from Bamber's (1969) classic twelve-letter pool (i.e., B, C, D, F, J, K, L, N, S, T, V, and Z) using JavaScript, maintaining the rules established by Martin (1978). In other words, no letter was repeated within the same list. The original study (Martin, 1978) did not specify which letters were used; thus, Bamber's set was selected for standardization. The word "READY" appeared on the screen, with an audible signal to indicate the start of the sequence in four seconds. The word "READY" then disappeared, and three pairs of letters were revealed at a rate of 0.5 seconds per pair of letters, while the other two pairs remained displayed horizontally from left to right. All pairs then disappeared simultaneously before the recall prompt was presented, following a standardized three-second interval that would facilitate consistent response timing and data coding. Participants were required to enter each letter into predefined boxes corresponding to the serial position within the sequence. In the original study, Martin (1978) used 72 sets; however, in this replication, we used 30 sets to equalize the length of this task with the digit span task, while maintaining a sufficient quantity for quality analysis.

Procedure

The experiment began with a welcome message, offering participants the option to complete the study in either French or English. Participants were informed that the study automatically entered full-screen mode and was not compatible with phones. In the event of technical problems, participants could contact the team for assistance. Participants completed a consent form, which informed them that their participation was voluntary and that they could withdraw from the study at any time. They also completed a socio-demographic questionnaire to obtain a descriptive representation of our sample, which represents a modification of Martin's (1978) study, where only the age range of participants was reported in the original research. The questions asked enabled us to obtain a sufficiently representative portrait of our sample (e.g., gender, age, education level, ethnicity). Participants were provided with general instructions to reduce variability in testing conditions. These included verifying speaker function, sitting an arm's length from their device, using assistive devices if needed, ensuring a quiet environment and avoiding interruptions during the study.

Instructions for the digit span task were followed, along with a short practice session to prevent confusion during the scored task. This practice was formatted identically to the main task and featured a three-digit example. Before each digit list, the word "Ready" was played audibly five seconds before the auditory presentation. Digits were presented at a rate of one per second. After the final digit, a text box appeared where participants typed the numbers they had just heard. They had up to 30 seconds to respond and could press a "Next" button to proceed. If no response was submitted after 30 seconds, the following list was displayed automatically. This protocol was repeated 30 times.

After an optional short break following the first task, instructions and a practice trial were then presented for the serial letter recall task, following the same format as the main trials. The six letters were presented as three successive pairs at a rate of two pairs per second. Each new pair appeared while the previous ones remained on the screen, and all disappeared together before the recall phase. Participants were instructed to enter the six letters in the exact order they were presented. Unlike the digit span task, this interface included six pre-arranged input boxes — three for each letter pair —since this task evaluated order accuracy rather than memory span. As in the previous task, participants had 30 seconds to respond or could press "Next" to continue. Each task lasted approximately 10 to 15 minutes, totaling 45 to 60 minutes to complete the study. At the end of the second task, participants were asked to self-evaluate their performance and attention throughout the study.

Scoring

For the digit span task, we followed the method used by Woodworth and Schlosberg (1954), as adapted by Martin (1978), which determines a participant's digit span by identifying the longest list length recalled twice, with partial credit awarded for correct performance on longer lists. This scoring system accounts for both consistency and upper-bound capacity by combining base span with a fractional value for sequences attempted beyond, assigning a base score with fractional increments for partial completions. Firstly, the base value is the maximum length of digit lists that the participant can recall correctly. For example, if a participant could correctly recall all lists of three to six digits, including the list of six, the base value would be six. Secondly, for the calculation of successes above the base value, we need to assess how often the participant succeeds in recalling lists of digits longer than the base value. Each success above the base value adds a third of a point to the total score. For example, if a participant correctly recalls all lists, including six digits, succeeds twice with a seven-digit list and once with a nine-digit list, the total score is calculated as follows: 6 (base value) + 2/3 (success with seven digits) + 1/3 (success with nine digits) = 7.

For the order recall, consistent with Martin's (1978) original study, two distinct scoring methods were used to evaluate the responses of the 79 participants: strict criteria and lenient criteria. The average performance of the participants was calculated using these two formats. For the strict recall condition, an item was considered correct if it was reported in the proper position, earning one point for each letter that matched. The total score was therefore the sum of the letters correctly memorized, with a maximum of 6 points if the entire sequence was recalled correctly. The lenient criteria, on the other hand, is more tolerant. An item was considered correct if it was inputted by the participant, regardless of the position in which it was reported. The total score is therefore the sum of correctly memorized items, with a maximum of 6 points, but this time the participant can earn points even for minor errors in letter order. This mode of coding acknowledges partial recall and more accurately reflects the participant's level of memory, even in cases of minor errors in order or position.

Statistical analysis

To carry out the statistical analyses and generate the figures, R program and Excel were used. Mean scores and standard deviations were calculated for the digit span task, as well as for the serial letter recall task, using both strict and lenient scoring criteria. Data was analyzed to evaluate correlations between performance on the digit span and serial letter recall tasks, following the approach used by Martin (1978). Three Pearson correlations were computed to assess the relationships between (1) digit span and strict order recall, (2) digit span and lenient order recall, and (3) strict and lenient order recall scores. Additionally, a split-half analysis, using the Spearman-Brown coefficient, was conducted to assess the reliability of the lenient correction in the ordered recall task. Finally, we conducted a t-test analysis to check whether there were significant differences in performance between our younger participants (aged 18–30) and our older participants (aged 31 and above). For all these analyses, statistical significance was set to $\alpha = 0.05$.

Results

The total sample ($N = 79$) had a mean digit span of 3.82 ($SD = 1.86$, range = 0 to 9), indicating substantial interindividual variability. The mean recall accuracy for the strict method was 0.62 ($SD = 0.25$) and 0.66 ($SD = 0.25$) for the lenient method. The strict and lenient recall measures were strongly correlated among all participants ($r = 0.99$, $p < 0.001$), indicating a near-linear relationship between the two measures. Both strict recall ($r = 0.33$, $p = 0.004$) and lenient recall ($r = 0.34$, $p = 0.002$) showed a moderate correlation with the digit span. A split-half reliability analysis was conducted on the lenient recall to confirm the internal consistency, since it was done in the original study. The Spearman-Brown coefficient was 0.80 ($p < 0.001$), supporting internal consistency.

The sample was divided into two groups to compare the results with those of the original study (see Figure 2). Group 1 ($N = 54$) consists of participants aged 17 to 30 years, with an average age of 21.78, which is comparable to the age group of the original study. Group 2 ($N = 25$) consists of participants aged 31 years or older, with an average age of 54.6. When analyzed by age group, Group 1 had higher recall scores for strict recall ($M = 0.69$, $SD = 0.2$) and lenient recall ($M = 0.73$, $SD = 0.21$) than Group 2 (strict: $M = 0.48$, $SD = 0.27$; lenient: $M = 0.52$, $SD = 0.27$). Group 1 had a similar score on the digit span ($M = 3.71$, $SD = 1.99$) compared to Group 2 ($M = 4.05$, $SD = 1.57$), with no significant differences. In Group 1, strict and lenient recall remained highly correlated ($r = .98$, $p < .001$), and both were moderately correlated with digit span (strict: $r = .55$, $p < .001$; lenient: $r = .58$, $p < .001$). The R^2 for the strict and lenient recalls were respectively 0.301 and 0.332; in other words, 30.1% of the variation in the strict recall can be explained by the variation in the digit span, and the variation in the digit span can explain 33.2% of the variation in the lenient recall. In older adults, strict and lenient recall were also highly correlated ($r = 0.99$, $p < 0.001$); however, no significant correlations were observed with digit span.

Independent sample T-tests were conducted to find significant differences between the two groups. The tests indicated no significant differences between the age groups in digit span. However, Group 1 outperformed Group 2 in both strict recall ($t(77) = 3.68$, $p < 0.001$) and lenient recall ($t(77) = 3.68$, $p < 0.001$).

Discussion

The results obtained in the current study of Martin's (1978) second experiment are consistent with the hypothesis that digit span is a reliable predictor of other memory tasks. However, our findings only partially align with those reported by Martin (1978). The results show significant, albeit weak, correlations between digit span scores and scores on the two serial recall tasks among all participants ($r = 0.32$ for strict recall and $r = 0.34$ for lenient recall; see Table 1), suggesting that individuals scoring higher on the digit span tend to perform better on immediate recall tasks. These results aligned with the original study to a greater extent among the group of young adults, with moderate correlations (Group 1) (strict: $r = 0.55$, $p < 0.001$, $R^2 = 0.301$; lenient: $r = 0.58$, $p < 0.001$, $R^2 = 0.332$). These results support the hypothesis that the digit span represents a contributory, but not exclusive, factor in working memory performance.

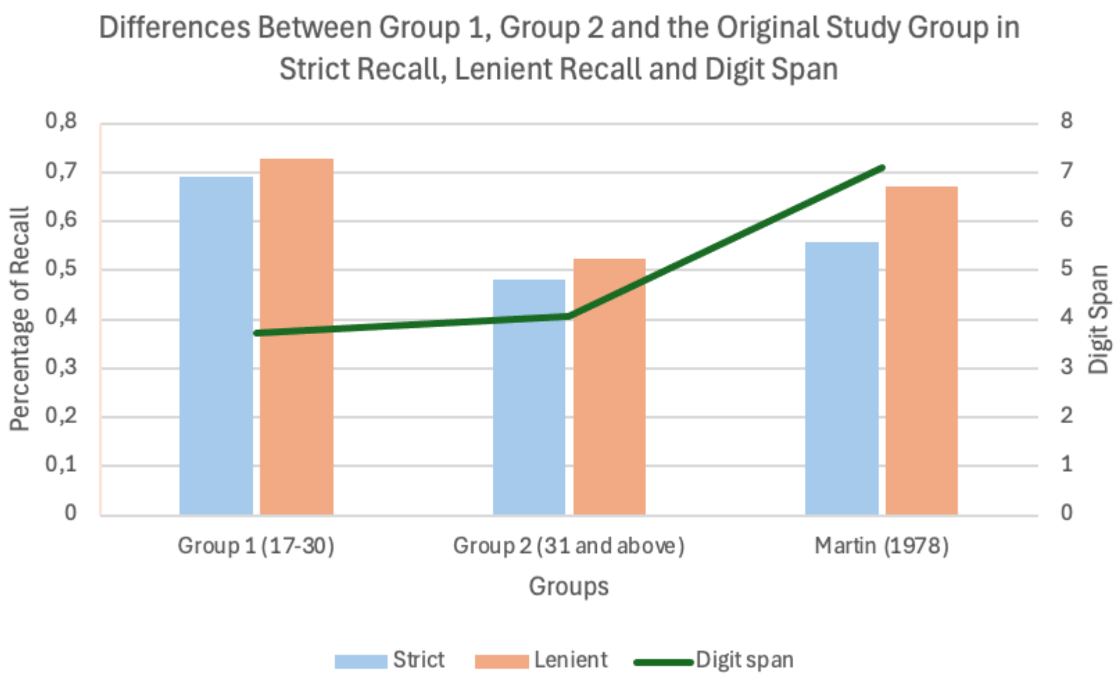
Furthermore, the similarity between the two correlations, in terms of statistical significance, suggests that the digit span does not discriminate between the order and the identity in which the items are recalled. Thus, the digit span could be a good and equally effective predictor for both types of information. These results partially confirm the findings of Martin (1978), whose original study revealed only significant correlations with the two serial recall methods (strict and lenient).

When we divided the sample into two age groups to allow comparison with the original study, we observed that the Group 1 (ages 17-30) scored higher on both the strict recall method ($M = 0.69$, $SD = 0.22$) and the lenient recall method ($M = 0.73$, $SD = 0.21$) compared to the Group 2 (ages 31 and above), which scored lower on strict recall ($M = 0.48$, $SD = 0.27$) and lenient recall ($M = 0.52$, $SD = 0.27$). The overall mean for the strict and lenient recalls were similar to the ones in the original study for the strict recall, with a mean of 0.62 ($SD = 0.25$) compared to 0.56 ($SD = 0.06$), and for the lenient recall, with a mean of 0.66 ($SD = 0.25$) compared to 0.67 ($SD = 0.05$; see Table 3). A surprising discovery was that Group 1 outperformed Group 2 by recalling both the items and their order using the two recall methods. In contrast, Group 2 achieved slightly higher, but not significant, digit span scores ($M = 4.05$, $SD = 1.57$) compared to Group 1 ($M = 3.71$, $SD = 1.99$). An interesting observation is that, for Group 1, both recalled tasks were significantly correlated with the digit span. In contrast, for Group 2, no such correlation was observed between the recall tasks and the digit span. This lack of correlation, alongside the discrepancy in scores between the ordered recall tasks and the digit span for Group 2, raises important questions regarding the underlying cognitive and working memory mechanisms. A recent study by Levi and Heled (2024) provides insight into these findings by examining age-related declines in working memory across different modalities. Their results showed that digit span performance was generally stronger than other modalities, indicating that verbal is the most skilled working-memory modality and tactile is the least skilled. Levi and Heled (2024) also reported an age effect, with younger adults (20-29 years) outperforming older adults (80-89 years). The weaker performance in the older group was linked to a general decline in working memory associated with age-related changes (Levi & Heled, 2024), which is attributed to alterations in neural structures in the frontal cortex along with more diffuse and attenuated neural activity.

Furthermore, their findings showed an interaction between age and modality. While the younger adults' group performed similarly across all modalities, the group of older adults showed only a significantly better performance on the digit span and a more preserved performance in the verbal modality. However, it exhibited declining performance in the visuospatial and, especially, tactile working-memory modalities. These findings may help to explain the discrepancy observed in the present study, as both ordered recall tasks were presented visually, whereas the digit span was presented verbally. Given that verbal information is more resilient to age-related changes, while visuospatial information declines more with age-related changes, the modality of presentation likely influenced the pattern of results (Levi & Heled, 2024).

However, the results of this study differ from those of the original study. The correlation between the digit span scores and the strict recall scores is weaker than that re-

Figure 2 ■ Differences Between the Groups of the Replication and the Original Martin (1978) Study. This figure illustrates the differences between Groups 1 and 2 from the replication study, in comparison to the original study group from Experiment 2 of Martin's (1978) study. The two tasks of the studies are illustrated, these being the digit span and order recall (i.e., strict and lenient correction) tasks. Strict criteria refer to correctly memorized letters, while lenient criteria refer to correctly inputting the letters regardless of the position in which it was reported.



ported in the original study, namely ($r = 0.63$, $p = 0.01$) in the original study, compared to ($r = 0.32$, $p = 0.004$) in this study. Furthermore, a significant increase in the correlation between digit span scores and lenient serial recall scores is observed ($r = 0.34$, $p = 0.002$). In contrast, the correlation was not significant in the original study ($r = 0.28$, $p > 0.05$). This difference may be attributed to the methodological strengths of the current study. We were able to design a reliable and standardized procedure, reducing the risk of human error while increasing internal validity, by utilizing current tools and technology. Additionally, these differences in correlation levels could be explained by the greater diversity and larger number of participants. This larger sample size may also explain why the Pearson correlation coefficient is significant in our study, compared to the one originally published in 1978. The Spearman-Brown coefficient (0.80, $p < 0.001$) was therefore not necessary, but still helpful in ensuring the internal validity of the items.

An unexpected result of our study is the sharp decrease in the average score on the digit span performances among current participants. While Martin's (1978) original study reported a mean of 7.10 ($SD = 1.18$), our study reveals a mean of 3.82 ($SD = 1.86$), which is consistent with citeapos(c01) conclusion that the limit of short-term memory is approximately 4 ± 1 items.

Table 1 ■ Correlations Among Measures of Replication Study.

Correlation	Total Sample ($N = 79$)	Group 1 (17-30)	Group 2 (31+)
Strict vs. Lenient Recall	$r = .99, p < .001$	$r = .98, p < .001$	$r = .99, p < .001$
Strict Recall vs. Digit Span	$r = .33, p = .004$	$r = .55, p < .001$	$r = .06$, ns
Lenient Recall vs. Digit Span	$r = .34, p = .002$	$r = .58, p < .001$	$r = .06$, ns

Note. ns means not statistically significant ($p > 0.05$)

Table 2 ■ Group Means and Standard Deviations of the Replication and the Original Martin 1978 Study

Measure	Total Sample (N=79)	Group1 (17-30)	Group 2 (31+)	Martin (1978)
Digit Span	3.82 (1.86)	3.71 (1.99)	4.05 (1.57)	7.1 (1.18)
Strict Recall	0.62 (0.25)	0.69 (0.22)	0.48 (0.27)	0.56 (0.06)
Lenient Recall	0.66 (0.25)	0.73 (0.21)	0.52 (0.27)	0.67 (0.05)

Nevertheless, these results reveal a nearly 50% drop in average performance compared to participants in the original study, indicating the presence of potential external influences on performance in our sample. Several studies have documented a slowdown, or even reversal, of the Flynn effect, which denotes an increase in population intelligence quotient (IQ) throughout the 20th century, suggesting a stagnation of IQ in certain populations (Bratsberg & Rogeberg, 2018). According to a study by Ritchie and Tucker-Drob (2018), social and technological changes over the past few decades have created cognitively impoverished environments, which may lead to a decline in cognitive abilities. Furthermore, a recent meta-analysis showed evidence that excessive screen-time exposure in children is associated with an increased risk of attention-related difficulties such as ADHD (Liu et al., 2024). Additionally, a subset of participants might have been presented with undiagnosed attention disorders, such as ADHD, which are known to impair working memory performance (Martinussen et al., 2005). Given the increasing prevalence of ADHD in recent decades (Xu et al., 2018), this trend may partially explain the lower digit span observed in the present study.

Strengths

This study represents the very first replication of Martin's (1978) study, as no other research has attempted to verify the possible presence of a Type 1 or Type 2 error. First, our sample size ($n = 79$) is significantly larger than that of the original study ($n = 16$), and the age range of our participants (17-74 years) is also much broader compared to that of the original study (18-30 years). This may provide a more accurate representation of the population, a more appropriate observation of the measured effect related to memory capacity, and more generalizable results. Another strength of this replication is the inclusion of a socio-demographic questionnaire, which provided a better characterization of participants and better control over certain individual variables, such as age, that can influence participants' performance and results. Furthermore, the integration of a break between the two tasks helped reduce the effects of fatigue and optimize participants' concentration. The use of modern tools and technologies contributed to minimizing variability in task administration and data collection, thereby ensuring greater reliability in the execution of the measures used, combined with standardized procedures, which strengthens the validity of the results obtained. Finally, although methodological adjustments were made, our replication remains faithful to the key elements of Martin's (1978) original study, ensuring scientific continuity, while adapting it to current standards in scientific research.

Limitations

Despite its strengths, our study has several limitations. Although the use of modern technology facilitated recruitment and protocol administration, it also represents a significant limitation. Participants completed the tasks online, in a semi-controlled environment, without supervision, which introduced variability in administration conditions (e.g., distractions, technical issues, distorted responses, unfamiliarity with technology). Unlike Martin's (1978) original study, our protocol did not ensure control over the testing environment, which may have influenced participants' performance. Furthermore, another limitation of our study is that it was not a fully faithful reproduction of the original study. The absence of counterbalancing the order of sequences using a Latin square may have led to practice or order effects. Finally, some participants may have used visual or spatial strategies, relying on personal sequences drawn from their own experiences, which could have influenced their performance.

Conclusion

This study aimed to replicate Martin's (1978) work to assess whether digit span constitutes a reliable predictor of individual differences in short-term memory, specifically in relation to order memory versus item memory. The results suggest significant correlations between digit span, and both ordered recall measures (strict and lenient), indicating that digit span can indeed reflect particular cognitive abilities, such as item memory and memory for their order of presentation. These findings highlight the limitations of the digit span test as a stand-alone measure of cognitive capacity. Furthermore, the ongoing debates about the actual capacity of working memory, ranging between 4 ± 1 items (Cowan, 2001) and 7 ± 2 (Miller, 1956), reinforce the idea that individual differences can only be understood through a multifaceted approach. Additionally, the modern methodology employed in this replication, including a more diverse sample and the use of digital technologies, may provide a more comprehensible overview of the potential generalizability of the results. However, it also introduces potential confounding factors, such as external distractions or technical difficulties. It is therefore essential to account for these variables in future studies by further standardizing experimental conditions. To summarize, this replication enriches the literature on working memory and underscores the importance of the ongoing empirical validation of measurement tools in cognitive psychology.

Suggestions

Regarding future research directions, it would be relevant to replicate Martin's (1978) study once again, this time incorporating a verbal span test that simultaneously assesses both item memory and memory for their order of presentation. It would also be worthwhile to examine the relationship between digit span and other types of memory beyond serial recall. A more in-depth exploration of cognitive factors, including individual differences such as age, general cognitive abilities, and demographic variables, could help better understand generational effects and contextual influences. Moreover, the use of varied stimuli (e.g., words, images) and pre-established, balanced stimulus lists could enhance the validity of the measurements. A controlled experimental environment would also help minimize distractions and ensure uniform testing conditions. Finally, the use of neuroimaging techniques, such as EEG and fMRI, could provide valuable insights into the brain activity involved in different phases of memory processing, thereby deepening our understanding of the mechanisms underlying memory span.

Authors' note

Various authors contributed to this article as part of an undergraduate course project. However, Cédric Champagne-Brassard, Gabrielle Tao, Laurie Rochon, and Vincent Poulin contributed an equal amount towards the preparation of this manuscript. Paola Michelle Garcia Mairena conceptualized, programmed, supervised, and assisted with the preparation of this manuscript.

References

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. doi: 10.1016/S1364-6613(00)01538-2.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, *8*(8), 47–89. doi: 10.1016/S0079-7421(08)60452-1.

Bamber, D. (1969). Reaction times and error rates for "same"-"different" judgments of multidimensional stimull. *Perception & Psychophysics*, *6*(3), 169–174.

Blankenship, A. B. (1938). Memory span: A review of the literature. *Psychological Bulletin*, *35*(1), 1–25. doi: 10.1037/h0061086.

Braden, J. P. (2024). Registered reports and replications: An ongoing journal of school psychology initiative. *Journal of School Psychology*, *103*, 101294–101294. doi: 10.1016/j.jsp.2024.101294.

Bratsberg, B., & Rogeberg, O. (2018). Flynn effect and its reversal are both environmentally caused. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(26), 6674–6678. doi: 10.1073/pnas.1718793115.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. doi: 10.3758/bf03196772.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. doi: 10.1017/s0140525x01003922.

Geva, S., Truneh, T., Seghier, M. L., Hope, T. M. H., Leff, A. P., Crinion, J. T., Gajardo-Vidal, A., Lorca-Puls, D. L., Green, D. W., & Price, C. J. (2021). Lesions that do or do not impair digit span: A study of 816 stroke survivors. *Brain Communications*, *3*(2). doi: 10.1093/braincomms/fcab031.

Giofrè, D., Stoppa, E., Ferioli, P., Pezzuti, L., & Cornoldi, C. (2016). Forward and backward digit span difficulties in children with specific learning disorder. *Journal of Clinical and Experimental Neuropsychology*, *38*(4), 478–486. doi: 10.1080/13803395.2015.1125454.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. *Psychology of Learning and Motivation*, *22*, 193–225. doi: 10.1016/s0079-7421(08)60041-9.

Huang-Pollock, C. L., & Karalunas, S. L. (2010). Working memory demands impair skill acquisition in children with adhd. *Journal of Abnormal Psychology*, *119*(1), 174–185. doi: 10.1037/a0017862.

Levi, O., & Heled, E. (2024). Aging processes of working memory in different modalities. *Neurology International*, *16*(5), 1122–1131. doi: 10.3390/neurolint16050084.

Liu, H., et al. (2024). Screen time and childhood attention deficit hyperactivity disorder: A meta-analysis. *Reviews on Environmental Health*, *39*(4), 643–50. doi: 10.1515/reveh-2022-0262.

Lynch, G., & Baudry, M. (2015). Brain and memory: Old arguments and new perspectives. *Brain Research*, *1621*, 1–4. doi: 10.1016/j.brainres.2014.12.052.

Majerus, S., Poncelet, M., Elsen, B., & van der Linden, M. (2006). Exploring the relationship between new word learning and short-term memory for serial order recall, item recall, and item recognition. *European Journal of Cognitive Psychology*, *18*(6), 848–873. doi: 10.1080/09541440500446476.

Martin, M. (1978). Memory span as a measure of individual differences in memory capacity. *Memory & Cognition*, *6*(2), 194–198. doi: 10.3758/bf03197446.

Martinussen, R., Hayden, J., Hogg-Johnson, S., & Tannock, R. (2005). A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *44*(4), 377–384. doi: 10.1097/01.chi.0000153228.72591.73.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? what does “failure to replicate” really mean? *American Psychologist*, *70*(6), 487–498. doi: 10.1037/a0039400.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. doi: 10.1037/h0043158.

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958. doi: 10.1037/bul0000153.

Rapport, M. D., Alderson, R. M., Kofler, M. J., Sarver, D. E., Bolden, J., & Sims, V. (2008). Working memory deficits in boys with attention-deficit/hyperactivity disorder (adhd): The contribution of central executive and subsystem processes. *Journal of Abnormal Child Psychology*, *36*(6), 825–837. doi: 10.1007/s10802-008-9215-y.

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? a meta-analysis. *Psychological Science*, *29*(8), 1358–1369. doi: 10.1177/0956797618774253.

Rosen, V. M., Bergeson, J. L., Putnam, K., Harwell, A., & Sunderland, T. (2002). Working memory and apolipoprotein e: What’s the connection? *Neuropsychologia*, *40*(13), 2226–2233. doi: 10.1016/S0028-3932(02)00132-X.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, *85*(3), 440–452. doi: 10.1037/0022-3514.85.3.440.

Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2011). Reliable digit span: A systematic review and cross-validation study. *Assessment*, *19*(1), 21–30. doi: 10.1177/1073191111428764.

Weiss, L. G., Saklofske, D. H., Holdnack, J. A., & Prifitera, A. (2016). Wisc-v: Advances in the assessment of intelligence. *WISC-V Assessment and Interpretation*, 3–23. doi: 10.1016/b978-0-12-404697-9.00001-7.

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*(433). doi: 10.3389/fpsyg.2013.00433.

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry*, *57*(11), 1336–1346. doi: 10.1016/j.biopsych.2005.02.006.

Woodworth, R. S., & Schlosberg, H. (1954). *Experimental psychology*. Oxford University Press.

Xu, G., Strathearn, L., Liu, B., Yang, B., & Bao, W. (2018). Twenty-year trends in diagnosed attention-deficit /hyperactivity disorder among us children and adolescents, 1997-2016. *JAMA network open*, *1*(4), e181471. doi: 10.1001/jamanetworkopen.2018.1471.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on the journal’s web site.

Citation

Garcia Mairena, P. M., Champagne, C., Poulin, V., Rochon, L., Tao, G., Allanach, J., Arenas, O., Arpaci, B., Attrassi, N., Berrada, Z., Bérubé, C., Bilodeau, C., Bouity, E., Boutin, A., Castonguay, M., Cléroux, M. P., Easton, É., Errachoud, S., Forbes, R., … Zaher, I. (2025). Exploring memory span and capacity: A replication of Martin (1978, Experiment 2). *The Quantitative Methods for Psychology*, *21*(3), 138–147. doi: 10.20982/tqmp.21.3.p138.

Received: 02/09/2025 ~ Accepted: 04/10/2025